Molecular Studies of Saline Antarctic Lakes from a Whole Ecosystem Perspective

Sheree Yau

2012

A thesis submitted for the degree of Doctor of Philosophy

School of Biotechnology and Biomolecular Sciences
University of New South Wales, Australia

Originality Statement

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published of written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.'

Signed .....

Date .....

Copyright Statement

Authenticity Statement

'I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis. No emendation of content has occurred and if there are any minor variations in formatting, they are the result of the conversion to digital format.'

Acknowledgements

List of Publications

Publications and submitted manuscripts arising from my PhD research are listed below. In all cases, my supervisor Prof Ricardo Cavicchioli and my co-supervisor Dr Federico Lauro were involved in the research design and editing of the manuscripts. Where versions of published material appears in this thesis, details of the contributions made by myself and others precede it.

- David Wilkins, **Sheree Yau**, Timothy Williams, Michelle Allen, Mark V. Brown, Matthew Z. DeMaere, Federico M. Lauro and Ricardo Cavicchioli. Key Microbial Drivers in Antarctic Aquatic Environments. *FEMS Microbiology Reviews* (in press), 2012.

- **Sheree Yau** and Ricardo Cavicchioli. Microbial communities in Antarctic lakes: Entirely new perspectives from metagenomics and metaproteomics. *Microbiology Australia* 32:157–159, 2011.

- Federico M. Lauro, Matthew Z. DeMaere, **Sheree Yau**, Mark V. Brown, Charmaine Ng, David Wilkins, Mark J. Raftery, John A.E. Gibson, Cynthia Andrews-Pfannkoch, Matthew Lewis, Jeffery M. Hoffman,Torsten Thomas and Ricardo Cavicchioli. An integrative study of a meromictic lake ecosystem in Antarctica. *ISME Journal* 5:879–895, 2011.

- **Sheree Yau**, Federico M. Lauro, Matthew Z. DeMaere, Mark V. Brown, Torsten Thomas, Mark J. Raftery, Cynthia Andrews-Pfannkoch, Matthew Lewis, Jeffery M. Hoffman, John A. Gibson and Ricardo Cavicchioli. Virophage control of antarctic algal host–virus dynamics. *Proceedings of the National Academy of Sciences USA* 108:6163–6168, 2011.

Abstract

Contents

List of Figures

List of Tables

List of Abbreviations

1D-SDS PAGE one dimensional-sodium dodecyl sulphate polyacrylamide gel electrophoresis

ABC ATP-binding cassette

BLAST basic local alignment search tool

CAMERA community cyberinfrastructure for advanced microbial ecology research and analysis

CAS CRISPR-associated proteins

COG clusters of orthologous groups

CRISPR clustered regularly interspaced short palindromic repeat

DGGE denaturing gradient gel electrophoresis

DMSO dimethylsulphoxide

DOC dissolved organic carbon

GOS global ocean sampling

GSB green sulphur bacteria

HMMER biosequence analysis using profile hidden Markov models

JCVI J.Craig Venter Institute

KEGG Kyoto Encyclopedia of Genes and Genomes

MS mass spectrometry

MS-MS two dimensional mass spectrometry

NCBI National Center for Biotechnology Information

NR non-redundant database

ORF open-reading frame

PCR polymerase chain-reaction

RNA ribonucleic acid

rRNA ribosomal RNA

SSU small subunit ribosomal RNA

SRB sulphate reducing bacteria

STAMP statistical analysis of metagenomic profiles

VLP virus-like particle

WGS whole genome shotgun

Chapter 1

General introduction: molecular microbial ecology of Antarctic lakes

Co-authorship statement

Sections of this chapter has been published as:

David Wilkins, **Sheree Yau**, Timothy Williams, Michelle Allen, Mark V. Brown, Matthew Z. DeMaere, Federico M. Lauro and Ricardo Cavicchioli. Key Microbial Drivers in Antarctic Aquatic Environments.
FEMS Microbiology Reviews (in press), 2012.

I contributed the section of the publication entitled *Antarctic lakes* excluding the subsection, *Microbial mats as microcosms of Antarctic life*. This material appears in section 1.6 *Molecular approaches used in Antarctic lake systems*, section 1.7 *Molecular insights into Antarctic lakes* and section 1.8 *limitations of taxonomic surveys* of this introduction.

1.1 Introduction

Antarctica is a "frozen desert" of constant low temperature, little precipitation and long polar light cycles where only specially adapted organisms can survive. The continent is covered by ice up to 4 km thick that covers an area of 13.8 million km^2. A tiny 0.32% of the land area is ice-free, most of which consists of exposed rocky peaks or nunataks such as in the Ellsworth, the Transantarctic and the North Victoria Land Mountains. Only 1–2% of that ice-free land comprises rocky coastal oases; however, it is these regions where Antarctic life is concentrated (Hodgson, 2012).

They are breeding sites for large animals such as seals, penguins and sea birds and some of the only locations where plants and lichens are found. Coastal oases are also distinguished by the presence of hundreds of lakes and ponds. Life in these lakes is microbially dominated with few, if any, metazoan inhabitants (Laybourn-Parry, 1997) making them ideal locations to study Antarctic microbiota. The lakes span a continuum of environmental factors such as salinity and are "natural laboratories" to examine adaptations to a property of interest. They are also ideal model ecosystems as they are largely isolated with a close relation between species and function.

This introduction will describe the Antarctic lakes, their microbiology and review molecular-based Antarctic microbiological research on the lakes. As this thesis focused on two lakes in the Vestfold Hills, emphasis will be given to describing research from this study site.

1.2 Antarctic lakes

In Antarctica, perenially available liquid water is found predominantly in lakes. These lakes span a wide range of physical and chemical properties from freshwater to hypersaline and constantly ice-covered to melted. Some are permanently stratified and termed meromictic if they thaw seasonally, or amictic if they are always ice-covered. Stratified lakes provide a unique opportunity to describe microbial populations along chemical gradients, but within a single water body. Most lakes are ice-covered for most of the year making them effectively isolated, and some may be truly closed systems if ice-cover is permanent. The age of water varies considerably; for example, outflow of subglacial water at Blood Falls is estimated to be 1.5 million years

old (Mikucki *et al.*, 2009) while water from Lake Miers is less than 300 years old (Green *et al.*, 1988). Overall, there are two main Antarctic lake types: those bound by ice, comprising subglacial; epiglacial and supraglacial lakes, and those bound by rock.

1.2.1 Ice-bound lakes

Subglacial lakes are pools of water beneath an ice sheet that form as the pressure of glacier flow against bedrock melts the ice at the interface. They are prevalent in Antarctica with at least 145 identified (Siegert *et al.*, 2005). The largest of these is Lake Vostok, which is 240 km long, 50 km wide and up to 1 km deep (Siegert *et al.*, 2001). Here the pressure is an extreme 340 atmospheres (Siegert *et al.*, 2001). These lakes are found dotted around the continent generally under the continental ice shelf (Siegert *et al.*, 2001).

Epiglacial lakes are similarly formed in the boundary between rock and ice, but where rock is exposed, such as where a glacier front contacts a mountain side. They are potentially the most common lake type as they can occur where ever there is rock and thus are both inland and coastal (Hodgson, 2012). These lakes can be highly changeable as they are subject to glacial movements and meltwater inputs. As a result, they can be short-lived, but many examples are thousands of years old, such as the .

Pockets of water can be also be found on top of glaciers. One example is cryoconite holes, which originate from the heat absorbed by dark dust melting small depressions on the glacier surface. These are extremely interesting systems with massive ranges in pH and chemistry. Most substantial supraglacial lakes also occur. However, all of the supraglacial reservoirs of water tend to be ephemeral lasting only during the summer months.

1.2.2 Rock-bound lakes

These lakes are of water trapped in exposed rocky basins. Mountain lakes.

By far the majority of lakes are found in the coastal oases. Most of these lakes were formed when the retreat of the continental ice-shelf lead to isostatic uplift of the land (Burton, 1981). As a result, the majority of lakes in the coastal oases are composed of relic seawater and are predominantly saline or hypersaline (Burke and Burton, 1988). In the latter, salinity is high due to concentrated by ablation (evaporation and sublimnation). Lakes closer

to the coastline may still occasionally experience marine inputs. Epishelf lakes are a type of lake unique to coastal regions.

Freshwater lakes near the continental ice shelf were likely already above sea-level as the ice receded and are not of marine origin (Bronge, 2004). Other freshwater lakes were originally marine-derived but have been flushed fresh by glacial meltwater (Pickard *et al.*, 1986). All lakes may receive water inputs from precipitation, from the ice-shelf and glacial melt streams (Burton, 1981). This can cause freshwater to seasonaly overlay some saline lakes as the ice-cover thaws. The chemistry of the exposed lakes is very much influenced by the water balance from local geographic and climatic conditions which leads them to have different physical and chemical properties.

1.3 Coastal oases

These are also the best studied systems as research stations are the most hospitable sites for research stations. Coastal oases, where lakes are found fringe the Antarctic continent. In East Antarctica these include the Vestfold Hills, Bunger Hills, Larsemann Hills, Syowa Oasis, Schirmacher Oasis, Grearson Hills and McMurdo Dry Valleys. In West Antarctic, the Peninsula, the sub-antarctic islands and maritime islands house multiple lakes. Of these locations, the best studied lake systems are those of the McMurdo Dry Valleys, The Vestfold Hills and the subantarctic islands.

1.4 The Vestfold Hills, East Antarctica

The Vestfold Hills is a ice-free region of approximately 400 km^2 on the eastern shore of the Prydz Bay, East Antarctica in the Australian Antarctic Territory (fig:vestfold map) (Gibson, 1999). The region is made up of three large peninsulae, Broad, Mule and Long Peninsula, separated by Fjords connected to the sea. Some of these are large, such as Ellis Fjord which is 10 km long, up to 100 m deep and has become a stratified system due to its restricted opening to the ocean (Burke and Burton, 1988). The region was formed approximately 10,000 years ago in the early Holocene from isostatic rebound (Burton, 1981).

The Vestfold Hills were first sighted and named in 1935 (Law, 1959). Only intermittent expeditions occurred in the area until the establishment of

Davis Station (68°33′S, 78°15′E) in 1957 (Law, 1959). A continuous presence has been maintained since. There Vestfold Hills region was immediately noted for its extensize ice-free land and the numerous lakes(Johnstone *et al.*, 1973).

The Australian Antarctic Data Centre lists more than 3,000 water bodies mapped in the Vestfold Hills, ranging in area from 1 to 8,757,944 m^2. More than 300 lakes and ponds have been described, including approximately 20% of the world's meromictic lakes (Gibson, 1999). These are of particular interest because the anoxic bottom waters help preserve a paleogeological record in the sediments. This can tell us about the region and particularly climatic changes. Stratified lakes provide a unique opportunity to describe microbial populations along chemical gradients, but within a single water body.

1.4.1 Biology of the Vestfold Hills

1.4.2 Lakes of the Vestfold Hills

Much early work was dedicated to the biology of the Vestfold Hills. What was interesting and special? What was the picture they had a microbial life?

1.4.3 History of studies of Organic Lake

1.4.4 History of studies of Ace Lake

1.5 Cultivation and microscopy-based Antarctic microbiology

Early microbiological surveys began X. Bacteria were detected by cultivation or by microscopy. Identification was limited to those species that could be isolated and appropriate identification tests performed. Eucarya were identified with microscopy based approaches.

1.5.1 Eucarya

1.5.2 Bacteria

1.5.3 Archaea

1.5.4 Viruses

As obligate parasites, culturing viruses is made problematic by the need to have a susceptible host in culture. Furthermore, host specificity can be extremely narrow so any assessment of viral diversity by cultivation is highly limited. This is compounded by the logistical constraints of conducting field work in the Antarctic.

Most studies of Antarctic viruses have been confined to diversity analyses based on the electron micrographs of virus-like particle (VLP) morphotypes or visibly infected cells. Electron micrographs are able to distinguish to some extent tailed bacteriophages such as myo- sipho- and podo- viruses from one another due to tail morphology. For tailess viruses with icosohedral symmetry, morphology alone provides hardly any distinguishing features apart from capsid size. Other metrics used to assess viruses in the environment include enumeration, calculated the virus to bacteria ratios, visibly infected cells and viral production rates.

Overall, pioneering work on viruses has made several noteworthy observations.

1. Viruses are possibly more abundant in high latitude lakes than lower latitude.
2. Burst sizes of viruses are lower in high latitude lakes than in low latitude.
3. Viral abundance appears to correlate positively with salinity.
4. As food chains in Antarctic Lakes are truncated, viruses may play an increased importance in Antarctic lakes, particularly in increasing secondary production through the microbial loop.
5.
6.

Cold environments are hypothesized to be a 'hotspot' of viral diversity. However, molecular methods are required to validate this claim.

1.6 Molecular approaches used in Antarctic lake systems

The majority of molecular-based studies of Antarctic aquatic microbial communities have made use of polymerase chain-reaction (PCR) amplification of small subunit ribosomal RNA (SSU) sequences to survey the diversity of *Bacteria* and in some cases *Archaea* and *Eucarya*. Microbial composition has been determined by cloning and sequencing of ribosomal RNA (rRNA) gene amplicons exclusively (Bowman *et al.*, 2000b,a; Gordon *et al.*, 2000; Christner *et al.*, 2001; Purdy *et al.*, 2003; Karr *et al.*, 2006; Matsuzaki *et al.*, 2006; Kurosawa *et al.*, 2010; Bielewicz *et al.*, 2011), although most studies have also made use of denaturing gradient gel electrophoresis (DGGE) to provide a molecular "fingerprint" of the community (**???????**Mikucki and Priscu, 2007; **?**; **?**; **?**). Functional genes have also been targeted using polymerase chain reaction (PCR) amplification to assess the potential of biochemical processes occurring, such as nitrogen fixation (**?**), ammonia oxidation (**?**), anoxygenic photosynthesis (**?**), and dissimilatory sulfite reduction (**?**Mikucki *et al.*, 2009).

1.7 Insights from Antarctic molecular studies

1.7.1 Bacterial diversity: adaptation to unique physical and chemical conditions

The vast majority of molecular studies of Antarctic lakes have focused on bacteria. Consistent with the wide range of physical and chemical properties of Antarctic lakes, a large variation in species assemblages have been found. While exchange of microorganisms must be able to occur between lakes that are in close vicinity to each other, the picture that has emerged from the data to date is that microbial populations are relatively unique to each type of isolated system. Nonetheless, certain trends in bacterial composition are also apparent.

Focusing on the similarities, lakes of equivalent salinities tend to have similar communities. Hypersaline lakes from the Vestfold Hills (Bowman *et al.*, 2000a) and McMurdo Dry Valleys (**??**) were all dominated by *Gammaproteobacteria* and members of the Bacteroidetes as well as harboring lower abundance populations of *Alphaproteobacteria*, *Actinobacteria*, and *Firmi-*

cutes. The surface waters of saline lakes resemble marine communities dominated by *Bacteroidetes*, *Alphaproteobacteria* and *Gammaproteobacteria*, but divisions such as Actinobacteria and specific clades of *Cyanobacteria* have been found to be overrepresented compared to the ocean (**?**). Sediments from saline lakes in the Vestfold Hills (Bowman *et al.*, 2000b) and Nuramake-Ike in the Syowa Oasis (Kurosawa *et al.*, 2010) were very similar, containing in addition to the surface clades, *Deltaproteobacteria*, *Planctomycetes*, *Spirochaetes*, *Chloroflexi* (green non-sulphur bacteria), *Verrucomicrobia* and representatives of candidate divisions. Plankton from freshwater lakes were characterized by an abundance of *Betaproteobacteria*, although *Actinobacteria*, *Bacteroidetes*, *Alphaproteobacteria* and *Cyanobacteria* were also prominent (**????**).

Bacterial diversity defined by nutrients

Differences in bacterial community structure are also influenced by nutrient availability. In studies of freshwater lakes in the Antarctic Peninsula and the South Shetland Islands, cluster analysis of DGGE profiles grouped together lakes of similar trophic status (**??**). Most of the variance in community structure could be explained by related chemical parameters such as phosphate and dissolved inorganic nitrogen. Similarly, three freshwater lakes, Moss, Sombre and Heywood on Signy Island are alike except that Heywood Lake is enriched by organic inputs from seals.

Bacterial composition in each lake changed from winter to summer and this was again correlated to variation in physico-chemical properties (**?**). The bacterial population of Heywood Lake had shifted from a dominance of *Cyanobacteria* towards a greater abundance of *Actinobacteria* and marine *Alphaproteobacteria* (**?**). This hints at a link between a copiotrophic lifestyle in the Heywood Lake *Actinobacteria* and inhibition of Antarctic freshwater *Cyanobacteria* by eutrophication. This type of study exemplifies how inferences can be made about taxa and function by examining population changes over time and over gradients of environmental parameters.

Bacterial biogeography

The relative isolation and diverse chemistries of the lakes facilitates biogeographical and biogeochemical studies. The anoxic and sulfidic bottom

waters of some meromictic lakes form due to a density gradient that precludes mixing. Although sedimentation from the upper aerobic waters may occur, there is little opportunity for interchange of species with the bottom water of lakes allowing for greater divergence in community composition as nutrients can become depleted and products of metabolism can accumulate. As a result, distinct distributions of bacterial groups can inhabit these strata, and different types of microorganisms can be found in equivalent strata in different lakes. A good example of this is the presence of common types of purple sulphur bacteria (*Chromatiales*)and green sulphur bacteria (GSB)(*Chlorobi*) in some meromictic lakes and stratified fjords in the Vestfold Hills (Burke and Burton, 1988), compared to diverse purple non-sulphur bacteria in Lake Fryxell in Victoria Land (**?**). In Lake Bonney, the east and west lobes harbor overlapping but distinct communities in the suboxic waters (**?**). The east lobe was dominated by *Gammaproteobacteria* and the west lobe by *Bacteroidetes*, illustrating how divergent communities can form from the same seed population. In contrast, ice communities are more readily dispersed by wind, aerosols and melt-water. 16S rRNA gene probes designed from bacteria trapped in the permanent ice-cover of Lake Bonney hybridized to microbial mat libraries sourced up to 15 km away (Gordon *et al.*, 2000). This demonstrates how a single lake may encompass microorganisms that are geographically dispersed, while also harboring others that have restricted niches and are under stronger selection pressure.

Bacterial diversity of Lake Vostok

Subglacial systems, such as Lake Vostok, have been isolated from the open environment for hundreds of thousands to millions of years (Siegert *et al.*, 2001). As a result they provide a reservoir of microorganisms that may have undergone significant evolutionary divergence from the same seed populations that were not isolated by the Antarctic ice cover. The uniqueness of these types of systems also creates a conundrum for studying them. Lake Vostok is approximately 4 km below the continental ice-sheet making it extremely difficult to determine suitable means for accessing the lake without inadvertently contaminating it with biological or chemical matter (**?????**). To date, molecular microbial studies have concentrated on the accretion ice above the ice-water interface (**??**). Accretion ice has been found to contain a low density of bacterial cells from *Alphaproteobacteria*, *Betaproteobacteria*,

Actinobacteria and *Bacteroidetes* divisions closely allied to other cold environments. Molecular signatures of a thermophilic *Hydrogenophilus* species were also identified in accretion ice raising the possibility that chemoautotrophic thermophiles were delivered to the accretion ice from hydrothermal areas in the lakes bedrock (**??**). However, interpretation of results from samples sourced from the Lake Vostok bore hole are very challenging as it is difficult to differentiate contaminants from native Vostok microorganisms. From a study that assessed possible contaminants present in hydrocarbon-based drilling fluid retrieved from the Vostok ice core bore hole, six phylotypes were designated as new contaminants (**?**). Two of these were *Sphingomonas* phylotypes essentially identical to those found in the accretion ice-core (**?**), which raises question about whether bacterial signatures identified from the ice-cores are representative of Lake Vostok water, and further highlights the ongoing problem of causing forward contamination into the lake.

1.7.2 *Archaea*: methanogens and haloarchaea

Archaea have been detected mainly in anoxic sediments and bottom waters from lakes that range in salinity from fresh to hypersaline, and those with known isolates are affiliated with methanogens or haloarchaea (Bowman *et al.*, 2000a,b; Purdy *et al.*, 2003; **?**; **?**). Anoxia allows for the growth of methanogenic *Archaea* that mineralize fermentation products such as acetate, H_2 and CO_2 into methane, thereby performing an important step in carbon cycling. The acetoclastic methanogens thrive in environments where alternative terminal electron acceptors such as sulfate and nitrate have been depleted.

One example of this is Lake Heywood where methanogenic *Archaea* were found to comprise 34% of the total microbial population in the freshwater sediment, the majority of which were *Methanosarcinales* which include acetate and C1-compound utilizing methanogens (Purdy *et al.*, 2003). Both H_2:CO_2 (*Methanogenium frigidum*) and methylamine/methanol (*Methanococcoides burtonii*) utilizing methanogens were isolated from Ace Lake (**??**) providing opportunities for genomic analyses (**??**) and a host of studies addressing molecular mechanisms of cold adaptation (e.g. (**??**).

In general, archaeal populations appear to be adapted to their specific lake environment. Sediments from saline lakes of the Vestfold Hills were inhabited by members of the *Euryarchaeota* typically found in sedi-

ment and marine environments with the phylotypes differing between the lakes examined (Bowman *et al.*, 2000b). While a phylotype similar to *Methanosarcina* was identified, the majority were highly divergent. Similarly, *Methanosarcina* and *Methanoculleus* were detected in Lake Fryxell but other members of the *Euryarchaeota* and *Crenarchaeota* (a single sequence) were divergent, clustering only with marine clones (Karr *et al.*, 2006). Based on the lake chemical gradients and the location of these novel phylotypes in the water column the authors speculated these archaea may be have alternative metabolisms such as anoxic methanotrophy or sulphur-utilization.

In sediments from Lake Nurume-Ike in the Langhovde region, 205 archaeal clones grouped into three phylotypes, with the predominant archaeal clone being related to a clone from Burton Lake in the Vestfold Hills, while the other two did not match to any cultivated species (**?**). In hypersaline lakes where bottom waters do not become completely anoxic, methanogens are not present and *Archaea* have extremely low abundance. For example, only two archeael clones of the same phylotype were recovered from deep water samples from Lake Bonney (**?**), and Organic Lake in the Vestfold Hills had an extremely low abundance of archaeal clones related to *Halobacteriales* (**?**). In contrast to these stratified hypersaline lakes, the microbial community in the extremely hypersaline Deep Lake is dominated by haloarchaea (**?**). Many of the clones identified from Deep Lake are similar to *Halorubrum* (formerly *Halobacterium*) *lacusprofundi* which was isolated from the lake (**?**).

1.7.3 *Eucarya* perform multiple ecosystem roles

Single-celled *Eucarya* are important members of Antarctic aquatic microbial communities. In many Antarctic systems, eucaryal algae are the main photosynthetic organisms and in others, only heterotrophic protists occupy the top trophic level. As eucaryal cells are generally large with characteristic morphologies, microscopic identifications have been used. However, microscopy is unable to classify smaller cells such as nanoflagellates with high resolution, although these may constitute a high proportion of algal biomass. For example, five morphotypes of *Chrysophyceae*, evident in Antarctic lakes were unidentifiable by light microscopy but were able to be classified using DGGE and DNA sequencing (**?**). Consistent with this, molecular studies specifically targeting eucaryal diversity (**??**Bielewicz *et al.*, 2011) have iden-

tified a much higher level of diversity than previously suspected, and the studies have discovered lineages not previously known to be present such as silicoflagellates (**?**) and fungi (**?**Bielewicz *et al.*, 2011).

Most eucarya in Antarctic lakes are photosynthetic microalgae that are present in marine environments with a wide distribution including chlorophytes, haptophytes, cryptophytes and bacillariophytes. Molecular methods have afforded deeper insight into the phylogenetic diversity within these broader divisions and have revealed some patterns in their distribution. Using 18S rRNA gene amplification and DGGE, the same chrysophyte phylotypes were identified in lakes from the Antarctic Peninsula and King George Island despite being 220 km apart (**?**) indicating these species may be well-adapted to Antarctica or highly dispersed. Similarly, an unknown stramenopile sequence was detected throughout the 18S rRNA clone libraries of Lake Bonney demonstrating a previously unrecognized taxon occupied the entire photic zone in the lake (Bielewicz *et al.*, 2011). In constrast, other groups showed distinct vertical and temporal distributions with cryptophytes dominating the surface, haptophytes the midwaters and chlorophytes the deeper layers during the summer while stramenopiles increased in the winter (Bielewicz *et al.*, 2011). Further studies are necessary to determine the basis for apparent specific adaptations of some species to particular lakes or lake strata, and for the cosmopolitan distribution of others. Here, molecular based research of the kind that has been applied to bacteria such as functional gene surveys will undoubtedly help answer these questions.

1.7.4 Functional gene studies of Antarctic lakes

1.7.5 Integrative studies to derive whole ecosystem function

1.8 Limitations of taxonomic surveys

Inferring functional potential from taxonomic surveys can be problematic due to species or strain level differences in otherwise related bacteria. For example, the majority of the *Gammaproteobacteria* in hypersaline lakes were relatives of *Marinobacter* suggesting that this genus is particularly adapted to hypersaline systems (Bowman *et al.*, 2000b; **?**; Matsuzaki *et al.*, 2006; **?**). Nonetheless, *Marinobacter* species from different lakes appeared biochemically distinct as isolates from hypersaline lake Suribati-Ike were all able to

respire dimethylsulphoxide (DMSO) but not nitrate (Matsuzaki *et al.*, 2006). In contrast, those from the west lobe of Lake Bonney were all able to respire nitrate (**?**). Interestingly, in the east lobe of the same lake, nitrate respiration was inhibited although a near-identical *Marinobacter* phylotype was present; it was speculated that the inhibition may have been caused by an as yet unidentified chemical factor (**??**).

This also applies to *Eucarya*, as the influence of flagellates on ecosystem function is not necessarily clear-cut as they can simultaneously inhabit several trophic levels. For instance, in Ace Lake the mixotrophic phytoflagellate *Pyramimonas gelidocola* derives a proportion of its carbon intake through bacterivory (**?**) but in the nearby Highway Lake, it uptakes dissolved organic carbon (**?**). This again illustrates potential limitations for deriving ecosystem level functions from taxonomic studies alone, even with taxa that appear physiologically straightforward.

1.9 '-omics' approaches

Metagenomic studies have assessed both the taxonomic composition and genetic potential of lake communities, and in some cases have linked function to specific members of the community (**?????**).When coupled with functional "omic" techniques (to date metaproteomics has been applied, but not metatranscriptomics or stable isotope probing), information has also been gained about the genetic complement that has been expressed by the resident populations (**?**).

1.9.1 Viruses

1.10 Objectives

This study aimed to use a primarily metagenomic approach complemented with microscopy and metaproteomics to gain an integrative understanding of whole ecosystems. Ace Lake and Organic Lake were chosen as the study sites because as meromictic systems, differences in the microbial population can be examined across vertical gradients, they are also largely isolated, as marine-derived systems, adaptation of marine microbiota to lake system can be examined, they are sites of moderate diversity and may be reservoirs of

unknown taxa. Using this methodology, not only can the taxonomic composition of the lakes be determined but also the functional potential of the microbial population and insight into the active members of the community. The objectives of the research were:

1. To develop complementary analyses to metagenomic analysis.
2. To determine the microbial and viral composition of the lake communities and their functional potential.
3. To identify the genes expressed by the community.
4. To reconstruct genomic information of dominant taxa of interest.
5. To integrate environmental and biological data to model the lake microbial interactions and geochemical processes.

Chapter 2

Metaproteogenomic analysis of Ace Lake

Co-authorship statement

Sections from this chapter 2 have been published as:

Federico M. Lauro, Matthew Z. DeMaere, **Sheree Yau**, Mark V. Brown, Charmaine Ng, David Wilkins, Mark J. Raftery, John A.E. Gibson, Cynthia Andrews-Pfannkoch, Matthew Lewis, Jeffery M. Hoffman,Torsten Thomas, and Ricardo Cavicchioli. An integrative study of a meromictic lake ecosystem in Antarctica. *International Society of Microbial Ecology Journal* 5:879–895, 2011.

I performed the metaproteomic mass spectra analysis, epifluorescence imaging, microbial and viral counts and wrote the corresponding sections of the publication. Only these parts of the publication are included in the results and discussion of this chapter.

Analyses performed by others that support the work presented in this chapter are as follows: Research was designed by Federico Lauro, Mark Brown, Torsten Thomas, John Gibson and Ricardo Cavicchioli. Sample collection was performed by Federico Lauro, Mark Brown, Torsten Thomas, Jeffery Hoffman and Ricardo Cavicchioli. DNA extraction and clone library preparation of 2006 samples was performed by Cynthia Andrews-Pfannkoch and Jeffery Hoffman of the J.Craig Venter Institute (JCVI). DNA sequenc-

ing quality control was performed by Matthew Lewis of the JCVI. Metagenomic sequence filtering, mosaic assembly and annotation was performed by Matthew DeMaere. Protein extraction, one-dimensional sodium dodecyl sulphate-polyacrylamide gel electrophoresis and liquid chromatography mass spectrometry performed by Charmaine Ng. Assistance in mass spectra analysis was provided by Mark Raftery.

Relation of this work to thesis objectives

2.1 Summary

2.2 Introduction

Ace Lake is a meromictic saline lake in the Vestfold Hills, Antarctica. It is the best studied of all the meromictic lakes in the Vestfold Hills and possibly Antarctica. Extensive physical, chemical and biological data has been collected from Ace Lake in the last decades.

Ace Lake is a highly stratified lake system, 25 m deep at its deepest point. It is ice-covered for approximately 9 months of the year and thaws some summers. Water is marine-derived and a largely neutral water balance has ensured salinity is close to that of seawater. Slightly fresher surface water of X can overlay the deeper waters when the ice-cover melts. The lake is physically separated into an aerobic mixolimnion, a steep chemocline/oxycline at 12.7 m and an anoxic monimolimnion below. The monimolimnion is sulfidic with accumulated methane. Initial diversity analysis of the sediment showed microbial diversity was reduced (Bowman *et al.*, 2000a).

As a physically and chemically well-characterised system of moderate diversity, Ace Lake was chosen as a model ecosystem upon which to implement an integrative study using high through-put metagenomic and metaproteomic analyses. Samples were obtained down the depth profile at 5, 11.5, 12.7, 18 and 23 m depths corresponding to each of the three zones. Sampling was conducted as part of the global ocean sampling (GOS) expedition (**?**) by using size fractionation of microbial biomass onto 3.0, 0.8 and 0.1 μm membrane filters. Microbial and viral diversity was assessed from the metagenomic dataset. Significant differences were found in taxonomic composition of each size fraction within each sample depth and stratification of the microbial community between the three zones of Ace Lake. The mixolimnion community is similar to a marine surface water assemblage consisting of a high abundance of the SAR11 clade of *Alphaproteobacteria* related to "*Candidatus* Pelagibacter ubique" and green algae of the order *Mamiellales* but of one order of magnitude reduced diversity (**?**). Unlike Southern Ocean surface water, the mixolimnion is overrepresented in *Cyanobacteria* related to*Synechococcus* and *Actinobacteria*, which may represent taxa that mark

the transition of a marine to lake community. A dense, near clonal population of green sulphur bacteria related to *Chlorobium* termed C-ace reside at the chemo/oxycline at 12.7 m (**??**). Below, in the anoxic monimolimnion is a diverse, primarily heterotrophic community with abundant sulphate reducing bacteria (SRB) and methanogenic *Archaea*.

Preliminary work on the metaproteome down the vertical profile of Ace Lake has been performed using the National Center for Biotechnology Information (NCBI) non-redundant database (NR) database. However, there were few protein identifications. Identification rate reduced as diversity of the sample increased. A focused metaproteogeonomic analysis conducted on the dense GSB layer using the matched metagenome as the database resulted in many more protein identifications compared to using the(**?**). Assessment of the genetic complement of Ace Lake showed a concurrent stratification of the functional potential in each zone of Ace Lake (**?**). Significantly, GSB appeared to be crucial in the lake ecosystem as they had the greatest genetic potential for nitrogen and carbon fixation as well as sulphur cycling(**??**). Metaproteomic analysis was able to identify which proteins were actively expressed and thus the active pathways of the GSB metabolism which are so crucial to the function of the lake (**?**).

This study aimed to expand on the metagenomic analysis of the water column of Ace Lake using complementary analyses. Metaproteomic analysis was used to identify expressed proteins of the 0.1 μm size fraction Ace Lake using a matched metagenomic databases for protein identification to infer which taxa and metabolic processes were active at time of sampling. To determine cellular and viral densities and validate the efficacy of size-fractionation with a modified epifluorescence microscopy procedure was developed and implemented.

2.3 Materials and methods

2.3.1 Ace Lake samples

Water samples were collected from Ace Lake (68°24′S, 78°11′E), Vestfold Hills, Antarctica on 21 and 22 December 2006. A 2 m hole positioned above the deepest point (25 m depth) of the lake was drilled through the ice cover of Ace Lake to reach the lake surface. A volume of 1–10 L was collected by sequential size fractionation through a 20 μm pre-filter directly onto filters

3.0, 0.8 and 0.1 μm pore-sized, 293 mm polyethersulfone membrane filters (Rusch et al., 2007), along the depth profile as described previously (Ng et al., 2010). Samples were taken in the order, 23, 18, 14, 12.7, 5 and finally, 11.5 m.

After samples from each depth were collected, the sample racks were sequentially washed with 2 $\times$ 25 L 0.1 N NaOH, 2 $\times$ 25 L 0.053% NaOCl, and 2 $\times$ 25 L fresh water. The sample hose was flushed with water from each depth before being applied to the filters. A *Chlorobium* signature was identified at 5 m, but not immediately above the GSB layer at 11.5 m. As the next sample taken after sampling at 12.7 m was at 5 m, and then 11.5 m, despite all equipment being thoroughly washed with bleach, sodium hydroxide and water, the simplest explanation for the GSB signature at 5 m is carry-over from sampling of the dense biomass at 12.7 m.

A sonde probe (YSI model 6600, YSI Inc., Yellow Springs, OH, USA) was used to record depth, dissolved oxygen content, pH, salinity, temperature and turbidity throughout the water column of the lake. Total organic carbon was determined using a total organic carbon analyzer, TOC-5000A (Shimadzu, Kyoto, Japan) equipped with a ASI-5000A auto sampler (Shimadzu), and particulate organic carbon by standard protocols (http://www.epa.gov/glnpo/lmmb/methods/about.html) at the Centre for Water and Waste Technology, UNSW.

2.3.2 DNA sequencing and data cleanup

DNA extraction and Sanger sequencing was performed on 3730xl capillary sequencers (Applied Biosystems, Carlsbad, CA, USA) and pyrosequencing on GS20 FLX Titanium (Roche, Branford, CT, USA) at the J.Craig Venter Institute in Rockville, MD, USA (Rusch et al., 2007). The scaffolds and annotations will be available via community cyberinfrastructure for advanced microbial ecology research and analysis (CAMERA) and public sequence repositories such as the NCBI and the reads will be available via the NCBI Trace Archive.

Sanger reads were trimmed according to quality clear ranges. The quality of pyrosequencing reads was assessed as follows: a Blast nucleotide database was created from the Sanger reads of the 0.1 μm fraction of samples GS230, GS231 and GS232. After blasting the corresponding pyrosequencing reads against each database with a minimum bitscore of 80 and maximum e-value

of 0.1, reads were binned according to length. The percentage of reads for each bin lacking a match to the Sanger read database was recorded. The percentage reads at least 25% repetitive after MDUST (Morgulis et al., 2006) analysis at default settings, and the percentage of reads containing Ns, were assessed. In contrast to earlier pyrosequencers (Huse et al., 2007), no length-dependent bias in reads containing Ns was observed. However, short reads had a disproportionately high number of repeats. Moreover, based on the proportion of reads with no match to the Sanger data set, both very short and very long reads had a disproportionately high number of errors; an observation that was previously reported (Huse et al., 2007).

On the basis of this analysis, a three step filtering process was applied to each sample. Reads were initially run through the Celera sffToCA (Miller et al., 2008) pre-processor followed by Lucy (Chou and Holmes, 2001) and finally, excluding the bottom 8% and top 3% of reads determined from the read length distribution. As the sffToCA (v5.3) pre-processor removes all reads with a perfect prefix of any other read it overcomes the 'perfect duplicates problem (Gomez-Alvarez et al., 2009). After this process, <5% of the reads belonged to clusters of duplicates with three or more reads, and clusters of orthologous groups (COG) of proteins classification of these reads showed an over-representation of category L (replication, recombination and repair) that includes mobile genetic elements, which are often duplicated, suggesting a potential biological significance for the duplicated reads. It is possible these residual duplications are a result of high gene copy number or localized fragility of DNA sequences that might be biasing the shear points.

2.3.3 Epifluorescence microscopy

Samples of unfiltered lake water and the flow-through from 3.0 and 0.8 μm filters from all depths were collected on November 2008 and fixed on site in formalin 1% (v/v). The samples were stored at -80°C for subsequent direct counts of cells and VLPs. Enumeration was performed according to the method of Patel et al. (2007) with modifications. Lake water samples were filtered onto 0.01 μm pore-size polycarbonate filters (25 mm Poretics, GE Osmonics, Minnetonka, MN,USA). Filters were air dried, then placed with the back of the filter on top of a 30 ml aliquot of 0.1% (w/v) molten low-gelling-point agarose and allowed to dry at 30°C. Samples were stained by the addition of 1 ml working solution (1/400 dilution in 0.02 μm filtered

sterile Milli-Q) of SYBR Gold (Molecular Probes, Eugene, OR, USA) to 25 ml of mounting medium (VECTASHIELD HardSet, Vector Laboratories, Burlingame, CA, USA). Stained samples were counted immediately, or stored at $-20°C$ for up to a week before counting. Samples were visualised under wide-blue filter set (excitation 460–495 nm, emission 510–550 nm) with an epifluorescence microscope (Olympus BX61, Hamburg, Germany).

2.3.4 Metaproteomic analysis

Proteins were extracted from membrane filters from all 0.1 μm fractions from the six depths (5, 11.5, 12.7, 14, 18 and 23 m), and one-dimensional sodium dodecyl sulphatepolyacrylamide gel electrophoresis one dimensional-sodium dodecyl sulphate polyacrylamide gel electrophoresis (1D-SDS PAGE) and in gel trypsin digestion, liquid chromatography and mass mass spectrometry (MS), and two dimensional mass spectrometry (MS-MS) data analysis and validation of protein identifications performed as previously described (Ng et al., 2010), with minor modifications. The specta generated were searched against the protein sequence database corresponding to that depth constructed from the 0.1 μm mosaic assemblies.

Mosaic assemblies were generated for each sample fraction using Celera whole genome shotgun (WGS) Assembler v5.3 (Myers et al., 2000). For each assembly, the runtime parameters used were as outlined for 454 sequencing data in the published standard operating procedure (http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title 14SFF_SOP). As none of the samples can be considered clonal, these are regarded as stringent assemblies (Rusch et al., 2007). Each 0.1 μm fraction assembly was a hybrid of Sanger and 454 read data, wherein the estimated genome size was manually set to minimize the number of unitigs from abundant organisms being falsely classified as degenerate (Rusch et al., 2007). Annotation of each sample fraction assembly was carried out using an in-house pipeline, wherein the pipeline stages consisted of genomic feature detection and subsequent annotation. Detected features consisted of open-reading frames (ORFs), transfer RNA and rRNA. Each detected ORF was further annotated by basic local alignment search tool (BLAST) comparison against NR, Swissprot and Kyoto Encyclopedia of Genes and Genomes (KEGG)-peptide sequence databases and by biosequence analysis using profile hidden Markov models (HMMER) comparison against TIGRFAM (Haft et al., 2001), COG

(Tatusov et al, 1997; Tatusov et al., 2003) and known marker genes (von Mering et al., 2007). In all cases the cut-off e-value was a maximum of 1e−5.

The number of protein sequences in each database were as follows: 5 m, 138,208; 11.5 m, 133,948; 12.7 m, 27,142; 14 m, 62,436; 18 m, 71,512; and 23 m, 128,878. Scaffold (version Scaffold_2_05_01, Proteome Software Inc., Portland, OR, USA) was used to validate MS/MS-based peptide and protein identifications. Peptide and protein identifications were accepted if they could be established at >95% and 99% probability, respectively, as specified by the Peptide Prophet algorithm (Keller et al., 2002). Protein identifications required the identification of at least two peptides.

Proteins that contained similar peptides and could not be differentiated based on MS-MS analysis alone were grouped to satisfy the principles of parsimony and are referred to as a protein group. Spectral counting was used to semi-quantitively estimate protein abundance. The total assigned spectra that matched to each identified protein were exported from Scaffold 2.0. For similar proteins that have shared peptides (a protein ambiguity group), spectra were assigned to the protein with the most unique spectra. To normalize for variation in total spectra acquired between sample replicates, the number of spectra of each protein was multiplied by the average total spectra divided by the total spectra of the individual replicate. The spectral count of each protein was averaged across the replicates. As longer proteins are more likely to be detected, the average spectral counts were divided by the length of the protein. This value is equivalent to the normalized spectral abundance factor (Florens et al., 2006; Zybailov et al., 2006). In order to compare the relative abundance of proteins between depths, the normalized spectral abundance factor was divided by the average read depth of the contig (scaffold or degenerate) to which the protein mapped.

If >90%of a scaffolds length consisted of surrogate (highly degenerate unitig) sequence, the average read depth of the surrogate was used. For identified proteins that were part of a protein group the longest protein length and largest read depth value in the group was used. Pairwise comparisons of each zone were conducted on COG assigned proteins. The normalized spectral counts from each protein was aggregated based on their COG annotation. All proteins that were part of an ambiguity group were confirmed to share the same COG annotation to ensure counts were not biased because of the common spectra.

The summed spectral counts from 5 and 11.5 m (mixolimnion), and 14, 18 and 23 m (monimolimnion) were pooled. Statistical significance of differences between each zone was assessed using Fishers exact test, with confidence intervals at 99% significance calculated by the NewcombeWilson method and HolmBonferroni correction (P-value cutoff of 1e−5) in statistical analysis of metagenomic profiles (STAMP) (Parks and Beiko, 2010). All proteins identified, including their gene identifier, normalized spectral abundance, COG and KEGG orthology identifiers, KEGG locus tag and matching COG or KEGG description are provided in Supplementary Table S1.

2.4 Results and discussion

2.4.1 Development of epifluorescence microscopy methodology

Motivation for visualising microbiota from size fractionated water samples was twofold. Firstly, visualising microbiota from water samples allows examination of cellular morphologies and enumeration of cells and VLPs. In particular, cellular and VLP densities are not obtainable from the metagenomic data, although relative abundances are, and necessitates an alternative method of determination. Secondly, size fractionation of suspended microbial biomass from aquatic environments has been utilised as part of the landmark Sargasso Sea metagenomic study (**?**) and GOS expedition (**?**). The process was intended to simplify the community compositions and focus on marine *Bacteria* and *Archaea*, which are normally of small cell size by examining only the 0.1 μm fraction. The Ace Lake samples were collected using the same sampling strategy as the GOS dataset but has sequence information from all three filter sizes. Thus, the visualisation of microbial morphologies from each stage of the filtration process helped to validate the filtration process.

It was necessary to develop a revised method for visualising cells and viruses with fluorescent nucleic acid dyes due to the discontinuation of ANODISC filters (Whatman) that have been long used for this purpose (**??**). Since conducting this research, a similar protocol has been developed in other labs, stressing the need that has grown in the field for this procedure. Clear polycarbonate filters were used instead with (see section 2.3 Materials and methods) with few modifications to the published methods which

ensures this procedure can be easily adopted for widespread use. The back ground fluorescence of the clear filters was low when stain is only incorporated into the mountant. The disadvantages to the ANODISC filters was that polycarbonate has a tendency to crinkle upon mounting so cells and VLPs are not on a single visual plane and they have no plastic edging for handling and so can be less robust. They also cannot be blotted dry as ANODISC filters can. To counteract this effect agarose was used to embed the filters to help flatten the membrane and aid in mounting. However, this was not strictly necessary if filters were dried well and the membrane mounted carefully so that it was pressed flat against the glass slide. Filtration onto very small pore-size also necessitated a very strong seal of the filter column against the glass. Development of fluorescence microscopy methodology using 0.01 μm pore-size polycarbonate filters for simultaneous cellular and viral counts shows:

1. Size fractionation procedure appeared effective.

2. Morphological differences supports stratification of the community.

3. Visualisation of the morphology supported the metagenomic data that saw size fractionated and taxonomically stratified community.

4. Virus to bacteria ratios tell us about the community. At 12.7m depth, the light levels, and the sharp transition in oxygen content and salinity (Fig. S2) favour the dominance of a very high-density (2.2×10^8 cells ml^{-1}) of a single type of GSB of the genus *Chlorobia*, refered to as C-Ace (Ng et al 2010). Viral signatures were essentially devoid in this zone. The ratio of bacteriophage to total viral population increased proportionally in the larger size fractions consistent with trophic analyses that indicate that the larger size fractions are mostly copiotrophic (Fig. S8) particle attached bacteria and therefore likely to be sensitive to lysogenic phage infection (Lauro et al 2009). The 23 m unfiltered lake water contained very high levels (1.3×10^8 VLPs ml^{-1}) of VLPs. The high diversity of bacteria and archaea in all size fractions of the monimolimnion (Fig. 2) is consistent with the presence of a high viral population (Rodriguez-Valera et al., NRM, 2009).

To be a viable alternative, validation of this method with viral samples of known densities needs to be performed. However, for the purposes of this study, which is to show the relative differences in morphotypes between sample depths and size fractions, this method was more than suitable for the purpose.

2.4.2 Development of metaproteomic methodology

1. Using a matched metagenome instead of NR for protein identification greatly increased the number of identifications. 1.1 Except at the bottom zone, likely because the community is too diverse so greater coverage of the metagenome is required. In parallel with taxonomic diversity increasing with depth (with the exception of the GSB layer), the rate of metaproteomic identification of proteins decreased with depth (Table S2). The majority of the proteins that were detected (e.g. 67% at 23 m) were for hypothetical proteins that tended to lack orthologs in well-characterized organisms, highlighting both the functional importance and novelty of this anaerobic zone of the lake.

2. More specific information could be assigned to the taxonomic groups such as. 2.1. The Actinobacteria sequences in the mixolimnion were associated with a diverse phylogenetic cluster (Luna cluster) mainly contributed by freshwater ultramicrobacteria (Hahn et al., 2003). Several Luna cluster isolates contain rhodopsin genes (Sharma et al., 2009) and similar gene sequences were present in the Ace Lake oxic zone data and found to be expressed (167820670 and 163154474; Table S2). 2.2.This is consistent with the identification of clustered regularly interspaced short palindromic repeat (CRISPR) associated CRISPR-associated proteins (CAS) proteins Cse2, Cse3 and Cse4 (165526330, 165526332 and 165526334, respectively) in the 12.7 m metaproteome (Table S2). The CAS gene locus (cas3, cse1, cse2, cse3, cse4, cas5, cas1b), to which the proteins map, shares its organisation with CAS loci of sequenced GSB, and groups with the *E. coli* subtype/variant 2. The CRISPR/CAS system is likely to confer phage resistance to C-Ace, akin to the role in other organisms (Karginov and Hannon 2010; Horvath and Barrangou 2010).

3. Using Scaffold to validate protein identification and perform spectral counts was helpful. 3.1Same protein identifications as Charmaine except one or two. 3.2 Able to quantify differences between mixolimnion and monimolimnion. The diversity and abundance of ATP-binding cassette (ABC) transporters was lowest in the 0.1 μm fractions at 23 m (Fig. 3), and a correspondingly low number were detected in the metaproteome (Table S2). In contrast, numerous transporters, predominately ABC type, were identified in the metaproteome of the mixolimnion samples, with a high COG representation of transporters for carbohydrates ($\approx$34% of normalized spectra),

amino acids (≈32%) and inorganic ions (≈9%) (Table S2 and Fig. S11). All transporters in the metaproteome were of bacterial origin and conservative phylum level assignments of the normalised spectra showed the majority to originate from *Proteobacteria* (69%), of which SAR11 comprised 46% and *Actinobacteria* 19% (Table S2). A high proportion of expressed genes with transport functions have also been reported for SAR11 from coastal (Poretsky et al. 2010) and open ocean waters (Sowell et al. 2009) (Morris et al. 2010?). Oligotrophs, such as SAR11 not only posses a low-diversity of high-affinity transporters (Lauro et al., 2009), but regulate the relative abundance of transporters expressed in response to dissolved organic carbon (DOC) availability (Poretsky et al. 2010). The prevalence of amino acid and simple sugar transporters (Table S2), and the low DOC concentration in the Ace Lake mixolimnion (Fig. 1) is likely to reflect efficient utilization of these substrates from the DOC pool. Two SAR11 transport proteins that were detected in Ace Lake (Table S2) were not detected from the Sargasso Sea (Sowell,et al. 2009): an ectoine/hydroxyectoine (167807477 and 167892279) and a zinc ABC transporter (167933120). The zinc ABC transporter is likely to support zinc efflux in response to zinc concentrations which are ∼70-fold higher in the mixolimnion of Ace Lake compared to seawater (Rankin 1999). Conversely, phosphate transporters were a major class detected from the Sargasso Sea (Sowell,et al. 2009) but were absent from the Ace Lake metaproteome; consistent with lower phosphate levels in the Sargasso Sea (<5 nM) compared to Ace Lake (1–12 μM). The differences in transporter expression between Ace Lake and oceanic SAR11 are likely to signify adaptive growth strategies that have evolved in the Ace Lake SAR11 community. The high numbers of bacteriophages in the monimolimnion (detected by microscopy, Fig. S5 and S6; metaproteomics, Table S2; metagenomics, Fig. 2), and increase in DOC observed at depth (Fig. 1), also indicates that carbon turnover in the monimolimnion is likely to be tightly coupled to the carbon flux going through a viral shunt, as proposed for open ocean systems (Suttle, C. A. Viruses in the sea. Nature 437, 356361 (2005)). The bacteriophages are also likely vehicles for mediating gene exchange. Most of the genetic potential to cycle the nitrogen pool appears to be limited to nitrogen assimilation throughout the lake and remineralization in the monimolimnion (Fig. S14). The detection of glutamine synthetase (GlnA) and glutamate synthases (GltBS) in the metaproteome (Table S2)

are supportive of active nitrogen assimilation. In the mixolimnion, GlnA was linked to SAR11 and Actinobacteria, and they are likely to be responsible for nitrogen absorption in the oxic zone. At the oxycline, GlnA and GltB from GSB were abundant (Table S2), indicating an important role for nitrogen assimilation at this zone in the lake. Genes for assimilatory sulphate reduction (ASR) were present in metagenome data of all three fractions at all depths, although they were lowest at the oxycline. However, there was no evidence for expression of the genes as ASR proteins were not detected by metaproteomics. In contrast, multiple subunits of the GSB dissimilatory sulfide reductase (DSR) complex were identified (Ng et al. 2010 and Table S2) indicating functionality of this pathway at the oxycline. GSB likely utilise the DSR system to convert sulphur to sulfite and the polysulfide-reductase-like complex 3 to oxidize sulfite to sulphate. SRB may then reform sulphide completing the sulphur cycle between the GSB and SRB (Ng et al. 2010). While SRB were detected at the three depths of the monimolimnion, sulphate is depleted in the water column and sediment at the bottom of the lake limiting their dissimilatory capacity (Rankin et al 1999). Finally, sulphate in the mixolimnion can be linked to sulphur-oxidation by SAR11 (Meyer and Kuever, Microbiology 153:3478-3498) and a concomitant lack of capacity to perform sulphur reduction.

2.5 Conclusions

Using complementary approaches helps to validate the research methodology and metagenomic inferences about the whole community. Specifically, differences in size and depth was shown by both microscopy and metagenomics to be apparent. This both validates the method of size fractionation as a viable approach to broad separation of the community, as well as supports the assertion that there was a large difference in community at different depths. Using a matched metaproteomic database matched to the metagenome showed a huge increase in the number of protein identifications. This was provided that metagenomic coverage was good. Using a metaproteomics, genes identified as potentially relevant in the metagenome were found to be expressed, supporting their importance. For example, it showed the CRISPR genes were active and may be a defence against phage. It also showed Actinorhodopsins were expressed. It showed that abundant

genes were normally abundant in the metaproteome, such as transport proteins. New inferences could be drawn from the metaproteome, such as the preference for labile substrates.

Chapter 3

Virophage control of Antarctic algal host–virus dynamics

Co-authorship statement

A version of this chapter has been published as:

Sheree Yau, Federico M. Lauro, Matthew Z. DeMaere, Mark V. Brown, Torsten Thomas, Mark J. Raftery, Cynthia Andrews-Pfannkoch, Matthew Lewis, Jeffery M. Hoffman, John A. Gibson, and Ricardo Cavicchioli. Virophage control of antarctic algal host–virus dynamics.
Proceedings of the National Academy of Sciences USA 108:6163–6168, 2011.

Contributions to this publication by other researchers is as follows. Research was designed by Federico Lauro, Mark Brown, Torsten Thomas, John Gibson and Ricardo Cavicchioli. Sample collection was performed by Federico Lauro, Mark Brown, Torsten Thomas, Jeffery Hoffman and Ricardo Cavicchioli. DNA extraction and clone library preparation of 2006 samples was performed by Cynthia Andrews-Pfannkoch and Jeffery Hoffman of the J. Craig Venter Institute. DNA sequencing quality control was performed by Matthew Lewis of the J. Craig Venter Institute. Metagenomic sequence filtering, global assembly and annotation was performed by Matthew DeMaere. Assistance in mass spectronomy and mass spectra analysis was pro-

vided by Mark Raftery. Assistance in analysis of Eucarya taxonomy was provided by Mark Brown. Analysis of virophage abundance over time was performed by Federico Lauro.

Apart from these contributions, I performed all other data analyses and interpretations.

3.1 Abstract

Viruses are abundant ubiquitous members of microbial communities, and in the marine environment affect population structure and nutrient cycling by infecting and lysing primary producers. Antarctic lakes are microbially dominated ecosystems supporting truncated food webs where viruses exert a major influence on the microbial loop. Here we report the discovery of a new virophage (relative of the recently described Sputnik virophage) that preys on phycodnaviruses that infect prasinophytes (phototrophic algae). By performing metaproteogenomic analysis on samples from Organic Lake, a hypersaline meromictic lake in Antarctica, complete virophage and near-complete phycodnavirus genomes were obtained. By introducing the virophage as an additional predator of a predator-prey dynamic model we determine that the virophage stimulates secondary production through the microbial loop by reducing overall mortality of the host and increasing the frequency of blooms during polar summer light periods. Virophages remained abundant in the lake two years later, and were represented by populations with a high level of major capsid protein sequence variation (25–100% identity). Virophage signatures were also found in neighbouring Ace Lake (in abundance), and in two tropical lakes (hypersaline and fresh), an estuary, and an ocean upwelling site. These findings indicate that virophages regulate host-virus interactions and influence overall carbon flux in Organic Lake, and play previously unrecognised roles in diverse aquatic ecosystems.

3.2 Introduction

It has been known for at least 20 years that viruses frequently infect and lyse marine primary producers causing up to 70% of cyanobacterial mortality (**??**). Eucaryotic phytoplankton are preyed upon by large dsDNA phycodnaviruses (PVs) causing bloom termination in globally distributed species (3,6). Elevated levels of dissolved organic carbon (DOC) (7) and numbers of heterotrophic bacteria (8-10) occur during algal blooms indicating that viral lysis of eucaryotic algae stimulates secondary production. Viruses also suppress host populations at concentrations below bloom-forming levels, with abundance being controlled by the efficiency and production rates of the infecting viruses (11, 12).

Antarctic lakes are microbially dominated ecosystems supporting few, if any metazoans in the water column (13, 14). In these truncated food webs, viruses are expected to play an increased role in the microbial loop (15). Low complexity Antarctic lake systems are amenable to whole community based molecular analyses where the role that viruses play in microbial dynamics can be unravelled (14). Attesting to this, a metagenomic study of Lake Limnopolar, West Antarctica uncovered a dominance of eucaryotic viruses and ssDNA viruses previously unknown in aquatic systems (16).

We established a metaproteogenomic program for Organic Lake (68°27′23.4″S, 78°11′22.6″E), which is located in the Vestfold Hills, East Antarctica, in order to functionally characterize its microbial community. Organic Lake is a shallow (7 m) hypersaline (≈230 g L^{-1} maximum salinity) meromictic lake with a high concentration of dimethylsulphide (≈120 μg^{-1}) in its anoxic monimolimnion (17, 18). Water temperature at the surface of the lake can vary from −14 to +15°C while remaining sub-zero at depth (19, 20). The lake is eutrophic, with organic material sourced both from autochthonous production and input from penguins and terrestrial algae. The high concentrations of organic material reflect slow breakdown in the highly saline lake water. The salt in the lake was trapped along with the marine biota when the lake was formed due to falling sea level c. 3,000 y BP (21, 22). The lake sediment has both low species diversity (Shannon-Weaver diversity: 1.01) and richness (Chao non-parametric index: 32 ± 12) (23). Unlike high latitude lakes, viral abundance has been reported to increase with trophic status (15) and with salinity in Antarctic lakes (24).

Here we report the analysis of the surface water of Organic Lake, highlighting the presence of a relative of the recently described Sputnik virophage, a small eucaryotic virus that requires a helper *Acanthamoeba polyphaga* mimivirus (APMV) to replicate (25). From metagenomic DNA, a complete Organic Lake virophage (OLV) genome was constructed (the second virophage genome to be described), and near-complete genomes of its probable helper Organic Lake phycodnaviruses (OLPVs).

3.3 Materials and methods

3.3.1 Samples and DNA sequencing

Water samples collected from Organic Lake were:

1. Surface water from the eastern side of the ice-free lake (68°27′25.48″S, 78°11′28.06″E) December 24, 2006.

2. A depth profile collected through a 30 cm hold drilled through the surface ice above the deepest point in the lake (68°27′22.15″S, 78°11′23.95″E), November 10, 2008.

3. surface water from the north-east side of the partially ice-covered lake (68°27′21.02″S 78°11′42.42″E), December 12, 2008.

Samples were sequentially filtered through a 20 μm pre-filter and biomass captured onto 3.0, 0.8 and 0.1 μm membrane filters as described previously (1, 2). The samples from 2008 also included 50% (v/v) RNAlater. DNA extraction, sequencing and quality validation was performed as previously described (1, 2). DNA sequencing was performed at the J. Craig Venter Institute in Rockville, MD, USA.

3.3.2 Transmission electron microscopy

Unfiltered Organic Lake surface water from December 24, 2006 (fixed on-site in 1% (v/v) formalin) was concentrated and a solvent exchange performed with sterile filtered ammonium acetate solution 1% (w/v) using a 50 kDa cut-off Microcon centrifugal filter device (Millipore) according to the manufacturers instructions. Formvar coated 200 mesh copper grids were floated on a droplet of sample for 30 min, excess liquid wicked off and the grid

negatively stained for 30 s with uranyl acetate 2% (w/v). The sample was visualised using a JEOL1400 transmission electron microscope at 100 kV at 150,000 to 250,000 × magnification.

3.3.3 Metagenomic assembly and annotation

Mosaic metagenomic assemblies were generated as previously described (1, 2). For the 0.1 μm Organic Lake 2006 sample, assembly was a hybrid of Sanger and 454 read data (Table S1). For all other sample size fractions, runtime parameters used were standard for 454 sequencing data. Low GC ($\geq$ 51%) scaffolds $>$ 10 kb from the 0.1 μm 2006 assembly had high coverage ($>$ 45 ×) indicating these were from the dominant taxa. One of these scaffolds was binned as virophage and the rest as PV.

To further separate the OLPV types and assess the completeness of their genomic content, highly conserved single copy PV orthologues were identified as follows. An all against all BLASTp search was conducted with protein sequences from the ten available PV genomes (*Acanthocystis turfacea Chlorella* virus 1, PbCV-1, PbCV AR158, PbCV FR483, PbCV NY2A, *Emiliania huxleyi* virus 86, *Ectocarpus siliculosus* virus 1, *Feldmannia* sp. virus, *Ostreococcus* virus 5, *Ostreococcus tauri* virus 1) and APMV (which was included as a close PV relative). BLASTp results were parsed and clustered using orthoMCL V1.4 (3, 4).

Pairs of each orthologue were located on eight of the PV scaffolds. The location of each orthologue pair had a complementary distribution so the eight scaffolds were able to be sorted unambiguously into two strains (OLPV-1 and OLPV-2). OLPV-1 ribonucleotide reductase α-subunit appeared as duplicated on different scaffold ends, likely as an artefact of its proximity to an assembly break point. The remaining high coverage scaffolds were searched for predicted proteins present in one OLPV strain but not in the other and assigned to the strain in which it was absent. Comparison of OLPV-1 and OLPV-2 scaffolds was performed using tBLASTn of concatenated scaffolds from each strain and visualised using the Artemis Comparison Tool (ACT) (5). DNA sequence data is available in Genbank and CAMERA (http://web.camera.calit2.net).

3.3.4 Organic Lake virophage genome completion and annotation

The high coverage (77 ×), large number of Sputnik homologues that encode essential functions and length of the putative OLV scaffold from the 0.1 μm 2006 hybrid assembly indicated it was a near-complete genome. Reads from this scaffold were reassembled at high stringency and visualised using Phred/Phrap/Consed (6) to complete the sequence. Mate-pair data indicated a circular molecule and primers were designed to span the ends of the scaffold and sequence across the gap (Table S5). Touch-down PCR was performed with eDNA from 0.1 μm 2006 sample, the product used for nested PCR and the final product was cloned and sequenced. The complete genome was manually annotated and visualised using Artemis (7). Translated ORFs (minimal size 120 amino acids) were compared (BLASTp) to GenBank, to the all metagenomic ORF peptide database on CAMERA (http://web.camera.calit2.net) and to predicted proteins from OLPV-1 and OLPV-2 scaffolds. Comparisons between the OLV genome and OLPV-1 /OLPV-2 were performed with tBLASTn and visualised using ACT (5).

3.3.5 Phylogenetic analysis

Translated amino acid sequences from viral marker genes of interest were retrieved from the 0.1 μm 2006 metagenomic assemblies from this study, GenBank and CAMERA all metagenomic reads ORF peptide database. Homologous sequences were aligned using MUSCLE v3.6 (8). Neighbour-joining analysis, test for clade support (bootstrap analysis 2000 replicates) and tree drawing was performed with Molecular Evolutionary Genetics Analysis (MEGA) software v4 (9). Maximum likelihood analysis (JTT substitution model) and test for clade support (aLRT analysis) was performed with PhyML (10) and the tree visualised using MEGA. 18S rRNA gene sequences were retrieved from reads of all filter sizes, compared (BLASTn, e-value $<$ 1.0e$-$5) to the SILVA100 SSURef database, aligned and phylogeny performed using ARB as previously described (1, 2). The abundance and similarity of virophages in all lake samples and filter sizes was estimated using BLASTp (evalue $<$ 1.0e$-$5) to search using the OLV MCP sequence against a database of proteins predicted from sequencing reads. The database was generated as previously described (1) and the percent

identity of the BLAST hit was used as a proxy for species similarity.

3.3.6 Metaproteomic analysis

Metaproteomics of proteins from the 0.1 μm filter from 2006 was performed as previously described (1, 2), with minor modifications. The protein sequence database was generated by combining ORFs from the 3.0, 0.8 and 0.1 μm mosaic assemblies with 130,581 sequences in the database. Scaffold 3.0 (Proteome Software Inc.) was used to validate MS/MS based peptide and protein identifications. Protein identification data is available in Table S2.

3.3.7 Model of algal host–virus and virophage dynamics

To model the effect a virophage would have on algal Pyramimonas algal host populations in Organic Lake, modified Lotka-Volterra equations were used describing the OLV as a predator of predator OLPV. The original equations are given by:

$$\frac{\mathrm{d}A}{\mathrm{d}t} = \alpha A - \varepsilon PA \tag{3.1}$$

$$\frac{\mathrm{d}P}{\mathrm{d}t} = \theta PA - \mu P \tag{3.2}$$

Where:

A is the number of *Pyramimonas* (prey).

P is the number of OLPV (predator).

α is the specific growth rate of the prey.

θ is the specific production rate of the predator.

ε is the rate of predator mediated death of prey.

μ is the specific decay rate of the predator.

Equation 3.1 describes the change in *Pyramimonas* abundance and equation 3.2 the change in OLPV abundance in the absence of OLV. In the presence of OLV, *Pyramimonas*, OLPV and OLV dynamics are described by the following equations:

$$\frac{\mathrm{d}P}{\mathrm{d}t} = \theta PA - \mu P - \omega PV \tag{3.3}$$

$$\frac{\mathrm{d}V}{\mathrm{d}t} = \beta PV - \gamma V \tag{3.4}$$

Where:

V is the number of teh OLV (predator of predator).

ω is the rate of OLV mediated reduction in OLPV infective particles.

β is the production rate of OLV.

γ is the decay rate of OLV.

Equation 3.3 is a modified version of equation 3.2 which includes the effect of OLV on the change in abundance of OLPV. Equation 3.4 describes the growth properties of OLV as a predator of OLPV. Values for the variables for the solution shown (Fig. 4) were as follows: initial prey (10), predator (1) and predator of predator (10) numbers, $\alpha = 0.1$, $\theta = 0.0015$, $\varepsilon = 0.01$, $\mu = 0.01$, $\omega = 0.01$, $\beta = 0.015$ and $\gamma = 0.15$. COmplex PAthway Simulator (COPASI) software (11) was used to simulate prey, predator and predator of predator dynamics using the deterministic (LSODA) method.

3.4 Results and discussion

3.4.1 Dominance of phycodnaviruses in Organic Lake

Water samples from Organic Lake were collected December 2006 and November and December 2008, and microbial biomass collected onto 3.0, 0.8 and 0.1 μm membrane filters as described previously (14). A large proportion of shotgun sequencing reads (96.2%) from the 0.1 μm size fraction of the 2006 Organic Lake metagenome (Table S1) had no significant hits to sequences in the RefSeq database (tBLASTx with e-value $< 1.0\mathrm{e}{-3}$, minimum alignment length: 60 bp, minimum identity: 60%). The degree of assembly was high, with 77% of reads forming part of a scaffold, indicating the sample contained a few abundant taxa of minimal diversity. Forty-five scaffolds were longer than 10 kb; the five longest ranged from 70 to 171 kb. GC content and coverage were used to separate scaffolds into taxonomic groups (Fig. S1). A broad division was evident between low ($\leq$ 41%) and high ($\geq$

51%) GC scaffolds suggesting they constituted two taxonomic groups. All scaffolds in the high GC group that could be assigned contained phage homologues, as did the one exceptional low GC scaffold. The low coverage in the high GC group showed bacteriophages were not abundant in the 0.1 μm fraction. These scaffolds were not analyzed further. The low GC scaffolds with confident assignments contained sequences matching conserved PV or APMV proteins. These PV-related scaffolds comprised 60% of assembled reads demonstrating that OLPVs were numerically dominant in the 0.1 μm fraction. Transmission electron microscopy (TEM) revealed the presence of virus-like particles with the dimensions and structure typical of PVs (Fig. 1A).

Within the low GC group, scaffolds separated into a high coverage ($>$ 45$\times$) group, including the five longest scaffolds, and a low coverage ($<$ 22$\times$) group. Two of the scaffolds in the high coverage group and one in the low coverage group contained the PV marker DNA polymerase B (DPOB). The two high coverage DPOB share 76% amino acid identity and both share $\approx$57% identity to the low coverage DPOB. DPOB is single-copy throughout the nucleo-cytoplasmic large DNA virus (NCLDV) family to which PVs belong (26,27), demonstrating that the Organic Lake surface waters contained two closely related abundant PV types (DPOB1) and (DPOB2), and a more distantly related lower abundance type (DPOB3).

Phylogenetic analysis clustered Organic Lake DPOB with unclassified lytic marine PV isolates that infect the prymnesiophytes *Chrysochromulina ericina* (CeV1) and *Phaeocystis pouchetii* (PpV), the prasinophyte *Pyramimonas orientalis* (PoV) (4,28), and uncultured marine PVs related to APMV (29, 30) (Fig. 2A and Fig. S2). As the host range of PVs broadly correlates with DPOB phylogeny (31, 32), OLPV would infect prasinophytes or prymnesiophytes. The most probable host is the prasinophyte, *Pyramimonas* (no prymnesiophyte 18S rRNA gene sequences were present in any size fraction of the Organic Lake metagenome) (Fig. S3).

Supporting the presence of more than one PV, pairs of single-copy PV orthologues (ribonucleotide reductase alpha and beta subunits, VV A32R virion packaging helicase, PBCV1 A482R-like putative transcription factor, VV D5 ATPase and VLTF2 family transcription factor) were identified in the high coverage scaffolds that shared an average of 81% percent amino acid identity. Based on the positions of single copy genes on the scaffolds and the

percent identity between them, the high coverage scaffolds were grouped into two strains designated OLPV-1 and OLPV-2 according to their DPOB phylogeny (Fig. 2A and Fig. S2). The remaining high coverage scaffolds were assigned to either strain, resulting in two near-complete genomes of ≈300 kb each (Fig. 2C), that are within the range of other sequenced PV genomes (155-407 kb). In addition, several OLPV genomic fragments contained PV homologues in high coverage scaffolds that could not be confidently assigned to either strain.

Both OLPV strains contain a PpV-like major capsid protein (MCP) designated MCP1 and another unique MCP designated MCP2 (Fig. 2B and Fig. S4). Both OLPV MCP1s were identified in the metaproteome (Fig. 2C and Table S2) but MCP2 was not. In addition to MCPs, the metaproteome contained a range of abundant structural proteins and others more likely to be packaged in the virion (e.g. chaperone), that were expressed by OLPV-1, OLPV-2 and/or an OLPV genomic fragment (Fig. 2C and Table S2). These data suggest that MCP1 is the major structural protein, and that both OLPV-1 and OLPV-2 were in a productive cycle in the lake at the time of sampling.

3.4.2 Complete genome of an Organic Lake virophage

Sputnik is a small (50 nm) icosahedral satellite virus of mamavirus (a new strain of APMV). It was termed a "virophage" because co-infection with Sputnik is deleterious to the mamavirus, resulting in abnormal virions and a decrease in mamavirus infectivity (25). One 28 kb scaffold in the low GC high coverage group had six out of 38 predicted proteins homologous to Sputnik virophage proteins (Fig. 3 and Table S3), and one PV homologue. The scaffold had a low GC content (≈30%), similar to the Sputnik genome, and was larger in size (28 kb vs 18 kb for Sputnik). Using PCR and sequencing, the scaffold was found to represent a complete circular virophage genome (the Sputnik genome is also circular). Virus-like particles resembling Sputnik in size and morphology were identified by TEM (Fig. 1B).

Sputnik homologues present in the Organic Lake scaffold included the V20 MCP, V3 DNA packaging ATPase, V13 putative DNA polymerase/primase and others of unknown function (V9, V18, V21 and V32) (Fig. 3 and Table S3). The Organic Lake virophage (OLV) is distinct to Sputnik as proteins

share 27-42% amino acid identity (28% MCP identity). OLV proteins include OLV9, the homologue of Sputnik V20 MCP, and OLV8, a fusion of the uncharacterised V18 and minor virion protein V19 from Sputnik (Fig. 3 and Table S3). The large number of homologues, including genes that fulfill essential functions in Sputnik (V20, V3 and V13), indicate that OLV and Sputnik have physiological similarities.

3.4.3 Gene exchange between virophage and phycodnaviruses

As PVs are related to APMV (27) and are abundant in Organic Lake, it stands to reason that OLPV is the helper of OLV. In the OLV genome, OLV12 is a *Chlorella* virus-derived gene, indicating that gene exchange has occurred between OLV and PVs (the function of OLV12 is discussed below). Similar observations were made for Sputnik, which carries four genes (V6, V7, V12 and V13) in common with the mamavirus, indicative of gene exchange between the viruses and possible co-evolution (25). As the V6, V7, V12 and V13 proteins have been associated with virophage-helper specificity, we reasoned that the functional analogues in OLV would have highest identity to proteins from its helper virus, rather than Sputnik.

By comparing OLV and OLPV, a 7,408 bp region was identified in OLV encoding five proteins (OLV17-22) with identity (32–65%) to sequences in both OLPV-1 and OLPV-2 (Fig. 2C, Fig. S5 and Table S3). OLV20 and OLV13 are collagen triple-helix-repeat-containing proteins, analogous to Sputnik collagen-like proteins (V6 and V7) involved in protein-protein interactions in the APMV virus factory. Sputnik can replicate with either mamavirus or APMV as a helper, although coinfection rates are higher with the mamavirus. V6 is the only protein with higher identity (69%) to mamavirus than APMV (42%) (25). Since OLV20 has equivalent identity (63%) with OLPV-1 and OLPV-2, it appears that OLV may be capable of interacting with both OLPV strains. Also within the conserved region, OLV22, is a 141 aa protein of unknown function that only matches sequences from OLPV and the Global Ocean Sampling (GOS) expedition (Table S3). Similar to OLV22, Sputnik V12 is a small protein (152 aa) of unknown function with high identity to APMV, and both may mediate a specific helper-virophage interaction. Other genes in this region of OLV can be mapped to OLVP, including a putative transmembrane protein (OLV17) and paralogous phage tail fibre repeat containing proteins, OLV18 and OLV19. Analogous to the

collagen-like proteins, OLV19 and OLV20 probably facilitate interactions between helper and virophage.

OLV12 (which is unique to OLV) consists of a C-terminal domain present in conserved hypothetical *Chlorella* virus proteins and an N-terminal domain most closely related to class 3 lipases that may confer OLV selectivity to a PV. OLV12 may function similarly to the Sputnik V15 membrane protein in modifying the APMV membrane (25). The Sputnik V13 consists of a primase domain and SF3 helicase domain related to NCLDV homologues, involved in DNA replication. The helicase domain of OLV25 and V13 are similar, although the primase domain is more similar to a protein from *Ostreococcus lucimarinus*, implying a past association of OLV with a prasinophyte alga host.

Genes unique to OLV point to adaptations specific to its helper-host system. Most notably, OLV possesses a N6 adenine-specific DNA methyltransferase, as does OLPV. In OLPV-1, genes for a bacterial type I restriction modification (RM) system are adjacent to a gene encoding a type I methylase-S target recognition domain protein, and upstream of a DNA helicase distantly related to type III restriction endonuclease (RE) subunits. A large number of *Chlorella* virus genomes have both 5mC and 6mA methylation (33), and several contain functional RM systems (34). The prototype *Chlorella* virus PbCV-1 possesses REs packaged in the virion for degrading host DNA soon after infection (35). In contrast to OLV/OLPV, DNA methyltransferases are absent in both Sputnik and APMV, indicating that the N6 adenine-specific DNA methyltransferase has been selected in OLV to reduce endonucleolytic attack mediated by OLPV.

3.4.4 Role of virophage in algal host–phycodnavirus dynamics

The presence of the virophage adds an additional consideration to the microbial loop dynamics. In batch amoeba cultures, co-infection of amoeba with APMV and Sputnik causes a 70% decrease in infective AMPV particles and a 3-fold decrease in lysis (25). To test how OLV affects OLPV and host population dynamics, we modelled the OLV as an additional predator of a predator in a Lotka-Volterra simulation (Fig. 4). In the model, the effect of virophage is robust, with equilibrium solutions across a wide range of parameter values (Fig. 4 shows one equilibrium solution). By decreasing

the number of infective OLPVs, the presence of OLV shortens the recovery time of the host population (Fig. 4C) and shifts the orbit away from the axis (Fig. 4D). The model reveals that the virophage stimulates the flux of secondary production through the microbial loop by reducing overall mortality of the host algal cell following a bloom, and by increasing the frequency of blooms during the summer light periods. Antarctic lake systems have evolved mechanisms to cope with long light-dark cycles (14) and shortened trophic chains. In Organic Lake and similar systems, a decrease in PV virulence may be instrumental in maintaining stability of the microbial food web.

3.4.5 Ecological relevance of virophages in aquatic systems

Metagenomic analysis of Organic Lake samples taken two years later in November (when the lake was ice covered) and December 2008 (partially ice-free) revealed sequences with 99% amino acid identity to OLV MCP indicating persistence of OLV in the ecosystem (Fig. 5 and Table S4). In addition, sequences with lower identity (25–90%) were detected, particularly in December, demonstrating Organic Lake virophages are highly diverse but OLV remained the dominant type.

From surface water samples of nearby Ace Lake (meromictic, surface 2% salinity), a large number of sequences were obtained that matched both the OLV MCP (Fig. 5 and Table S4) and PVs (14). All Ace Lake size fractions contained matches to OLV MCP, some with high identity (80–100%) and the majority with greater variation (25–80% identity) (Fig. 5 and Table S4). In contrast to Organic Lake where the largest number of matches was to the 0.1 μm size fraction, the majority of Ace Lake sequences were from the larger fractions (Fig. 5 and Table S4). This indicates the Ace Lake virophages were associated with host cells during sampling, or possibly with helper viruses that are larger than the OLPVs.

Extending the OLV MCP search to the GOS data revealed matches (25-28% identity) to sequences from the hypersaline Punta Cormorant Lagoon (Floreana Island, Galapagos), an oceanic upwelling near Fernandina Island (Galapagos), Delaware Bay estuary (NJ, USA), and freshwater Lake Gatun (Panama) (Table S4). The phylogenetic analysis of a conserved 103 amino acid region of the MCPs revealed a number of clusters, with Sputnik clustering with virophage sequences from Ace Lake that had low identity (22%)

to OLV MCP (Fig. 5 and Fig. S4). To improve searches for virophages and better understand their physiology and evolution, it will be valuable to target more genomes (e.g. the Ace Lake 167858124 relative with 40% MCP identity to Sputnik) and determine which genes are core to virophages and what relationship exists between genome complement and MCP identity.

In view of the implications of the virophage modelling (Fig. 4), the abundance and persistence of OLV in Organic Lake (Fig. 5, Table S4), and the presence of diverse virophage signatures in a variety of lake systems (fresh to hypersaline), an estuary, an ocean upwelling site and a water cooling tower (Sputnik), our study indicates that numerous types of virophages exist and play a previously unrecognised role in regulating host–virus interactions and influencing ecosystem function in aquatic environments.

3.5 Acknowledgements

We thank Craig Venter, John Bowman, Louise (Cromer) Newman, Anthony Hull, John Rich and Martin Riddle for providing helpful discussion and logistical support associated with the Antarctic expedition, and Lisa Ziegler for discussion about marine viruses. We acknowledge technical support for computing infrastructure and software development from Intersect, and in particular assistance from Joachim Mai. This work was supported by the Australian Research Council and the Australian Antarctic Division. Funding for sequencing was provided by the Gordon and Betty Moore Foundation to the J. Craig Venter Institute. Mass spectrometric results were obtained at the Bioanalytical Mass Spectrometry Facility within the Analytical Centre of the University of New South Wales. This work was undertaken using infrastructure provided by NSW Government co-investment in the National Collaborative Research Infrastructure Scheme. Subsidized access to this facility is gratefully acknowledged. We thank Jenny Norman from the UNSW Electron Microscopy Unit her assistance in generating images.

Chapter 4

Globally important biogeochemical processes from a hypersaline Antarctic lake

Co-authorship Statement

Chapter 5

Ace Lake Viral Genomes

Chapter 6

Conclusions and future work

Some ideas for this section include

Perspective of Antarctic Lake research from wetlab to molecular age. Summary of all molecular work done by our group to present. Summary of the major achievements of my work.

Needed future work for virophages. Since publication of my work, more virophages have been found. Need to isolate and track them over a season. Determine which OLPV type it infects. Verify OLPV infects pyramimonas. Verify OLV reduces infective particles. Make an exclusion experiment to show that the dynamics change with and without the OLV.

References

Bielewicz S., Bell E., Kong W., Friedberg I., Priscu J. C., and Morgan-Kiss R. M. Protist diversity in a permanently ice-covered Antarctic lake during the polar night transition. *The ISME journal*, 5(9):1559–1564, 2011.

Bowman J. P., McCammon S. A., Rea S. M., and McMeekin T. A. The microbial composition of three limnologically disparate hypersaline Antarctic lakes. *FEMS Microbiology Letters*, 183(1):81–88, 2000a.

Bowman J. P., Rea S. M., McCammon S. A., and McMeekin T. A. Diversity and community structure within anoxic sediment from marine salinity meromictic lakes and a coastal meromictic marine basin, Vestfold Hills, Eastern Antarctica. *Environmental Microbiology*, 2(2):227–237, 2000b.

Bronge C. Hydrographic and climatic changes influencing the proglacial Druzhby drainage system, Vestfold Hills, Antarctica. *Antarctic Science*, 8(4):379–388, 2004.

Burke C. and Burton H. R. Photosynthetic bacteria in meromictic lakes and stratified fjords of the Vestfold Hills, Antarctica. *Hydrobiologia*, 165(1): 13–23, 1988.

Burton H. R. Chemistry , physics and evolution of Antarctic saline lakes. *Hydrobiologia*, 82(1):339–362, 1981.

Christner B. C., Mosley-Thompson E., Thompson L. G., and Reeve J. N. Isolation of bacteria and 16S rDNAs from Lake Vostok accretion ice. *Environmental Microbiology*, 3(9):570–577, 2001.

Gibson J. A. The meromictic lakes and stratified marine basins of the Vestfold Hills, East Antarctica. *Antarctic Science*, 11(2):175–192, 1999.

Gordon D., Priscu J. C., and Giovanonni S. J. Origin and Phylogeny of Microbes Living in Permanent Antarctic Lake Ice. *Microbial ecology*, 39 (3):197–202, 2000.

Green W. J., Angle M. P., and Chave K. E. The geochemistry of Antarctic streams and their role in the evolution of four lakes of the McMurdo Dry Valleys. *Geochimica et Cosmochimica Acta*, 52(5):1265–1274, 1988.

Hodgson D. A. Antarctic Lakes. In Bengtsson L., Herschy R. W., and Fairbridge R. W., editors, *Encyclopedia of Lakes and Reservoirs*, pages 26–31. Springer, Dordrecht, 2012.

Johnstone G., Brown D., and Lugg D. The biology of the Vestfold Hills, Antarctica. *ANARE Scientific Reports*, 123:1–60, 1973.

Karr E. A., Ng J. M., Belchik S. M., Matthew W., Madigan M. T., Achenbach L. A., and Sattley W. M. Biodiversity of Methanogenic and Other Archaea in the Permanently Frozen Lake Fryxell , Antarctica. 72(2): 1663–1666, 2006.

Kurosawa N., Sato S., Kawarabayasi Y., Imura S., and Naganuma T. Archaeal and bacterial community structures in the anoxic sediment of Antarctic meromictic lake Nurume-Ike. *Polar Science*, 4(2):421–429, 2010.

Law P. The Vestfold Hills. *ANARE Reports*, 1:1–50, 1959.

Laybourn-Parry J. The microbial loop in Antarcti lakes. In Howard-Willsiams C., Lyons W., and Hawes I., editors, *Ecosystem Dynamics in a Antarctic Ice-Free Landscapes*, pages 231–240. Rotterdam, 1997.

Matsuzaki M., Kubota K., Satoh T., Kunugi M., Ban S., and Imura S. Dimethyl sulfoxide-respiring bacteria in Suribati Ike, a hypersaline lake, in Antarctica and the marine environment. *Polar Bioscience*, 20:73–81, 2006.

Mikucki J. A. and Priscu J. C. Bacterial diversity associated with Blood Falls, a subglacial outflow from the Taylor Glacier, Antarctica. *Applied and environmental microbiology*, 73(12):4029–39, 2007.

Mikucki J. A., Pearson A., Johnston D. T., Turchyn A. V., Farquhar J., Schrag D. P., Anbar A. D., Priscu J. C., and Lee P. A. A contemporary microbially maintained subglacial ferrous "ocean". *Science (New York, N.Y.)*, 324(5925):397–400, 2009.

Pickard J., Adamson D. A., and Heath C. W. The evolution of Watts Lake, Vestfold Hills, East Antarctica, from marine inlet to freshwater lake. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 53:271–288, 1986.

Purdy K., Nedwell D., and Embley T. Analysis of the sulfate-reducing bacterial and methanogenic archaeal populations in contrasting Antarctic sediments. *Applied and environmental ...*, 69(6):3181–3191, 2003.

Siegert M. J., Ellis-Evans J. C., Tranter M., Mayer C., Petit J.-R., Salamatin A., and Priscu J. C. Physical, chemical and biological processes in Lake Vostok and other Antarctic subglacial lakes. *Nature*, 414(6864):603–609, 2001.

Siegert M. J., Carter S., Tabacco I., Popov S., and Blankenship D. D. A revised inventory of Antarctic subglacial lakes. *Antarctic Science*, 17(03): 453–460, 2005.

Chapter 7

Appendices